Nutrition-Aware Food Image Classification Using Vision Transformers on Food-101

Yuze Jin (54587468)
STATS 507 – Data Science Analytics using Python

***Abstract*—Accurate dietary tracking plays an essential role in long-term health management, yet many nutrition applications still require users to manually enter food information. This process is time-consuming and error-prone, often resulting in incomplete or inconsistent dietary logs. Image-based food recognition offers a promising alternative due to the ubiquity of smartphone cameras. However, food classification remains challenging because of extreme visual variability, ambiguous category boundaries, and cluttered backgrounds.**

In this project, I develop an end-to-end nutrition-aware food recognition system using the Food-101 dataset and a pretrained Vision Transformer (ViT-Base/16). I implement a complete training pipeline using Hugging Face `datasets` and `transformers`, perform fine-tuning with data augmentation, evaluate model performance, and analyze misclassification patterns. The resulting model achieves 87.05% Top-1 and 97.50% Top-5 accuracy on the official test split ($n = 25{,}250$). I further incorporate a lightweight nutritional lookup module based on USDA FoodData Central, demonstrating how predicted classes can be mapped to approximate macronutrient estimates. Confusion-pair visualizations indicate that most errors arise from subtle inter-class similarities or multi-item scenes. Overall, these findings highlight both the strengths and limitations of transformer-based models in nutrition-oriented computer vision applications.

***Index Terms*—Food image classification, Vision Transformer, Food-101, Hugging Face, nutrition estimation.**

I. Introduction

Maintaining an accurate food diary is one of the most effective strategies for long-term health management. However, manually logging food names, ingredients, and serving sizes is burdensome and often inconsistent. Because taking a photo requires minimal user effort, image-based food recognition has become an appealing alternative for simplifying diet tracking.

Food classification, however, presents significant challenges. Dishes exhibit substantial intra-class variation depending on preparation style, lighting, viewing angle, and plating. Conversely, visually distinct dishes may share similar global appearance. Many images also contain multiple food items, yet datasets such as Food-101 assign only a single label. These difficulties require models capable of strong generalization and robust feature extraction.

Traditional convolutional neural networks (CNNs) [2] have shown strong performance on food recognition tasks. More recently, the Vision Transformer (ViT) [3] has emerged as a powerful alternative, leveraging global self-attention to capture spatially distributed discriminative cues—a property especially relevant for food imagery.

Beyond classification, nutrition-oriented applications require connecting predicted food categories to nutrient information. Although precise nutrient estimation from images remains extremely challenging, mapping predicted classes to representative nutrient profiles from standardized databases such as USDA FoodData Central [4] enables an interpretable proof of concept.

This project focuses on three main objectives:

- Implement and fine-tune a ViT-Base/16 model on the Food-101 dataset using a reproducible Hugging Face workflow.
- Evaluate classification accuracy, analyze misclassification patterns, and visualize frequent confusion pairs.
- Demonstrate a nutrition-aware pipeline by linking model predictions to USDA nutrient profiles.

II. Method

A. Problem Formulation

Let $x \in \mathbb{R}^{224\times224\times3}$ denote an RGB food image and $y \in \{1,\dots,101\}$ its class label. A model f_θ outputs class probabilities $\hat{p}$ via a softmax layer. Training minimizes the cross-entropy loss,

$$\mathcal{L}(\theta) = -\log \hat{p}_y. \tag{1}$$

At inference, the predicted class is $\hat{y} = \arg\max_i \hat{p}_i$, which is subsequently mapped to a nutrient profile.

B. Dataset and Splits

Food-101 [1] contains 101 food categories with 1,000 images each. Using Hugging Face `datasets`, I:

- load the dataset and use the official test split for final evaluation,
- create a 90/10 stratified validation split from the training portion using seed 507.

C. Preprocessing and Augmentation

Images are converted to RGB and resized to 224×224. Data augmentation is applied only during training and includes:

- random resized cropping,
- horizontal flipping,
- mild color jitter,
- small-angle rotation.

All images are then normalized using ImageNet statistics.

D. Vision Transformer Fine-Tuning

I fine-tune the `google/vit-base-patch16-224-in21k` model, which partitions each image into 16×16 patches and processes them via multi-head self-attention. The training configuration is:

- learning rate: 5×10^{-5},
- batch size: 16,
- weight decay: 0.05,
- training epochs: 1,
- mixed precision: `fp16`.

Despite the short training schedule, ViT models typically transfer effectively to downstream tasks.

E. Nutrition Mapping

To provide nutrition-aware output, I construct a lookup table linking each Food-101 class to representative USDA nutrient values, including calories and macronutrients per 100g. Although approximate, this demonstrates how classification predictions can support lightweight nutrition estimation.

III. Results

A. Accuracy and Evaluation

The final model achieves:

- **Top-1 accuracy: 87.05%**,
- **Top-5 accuracy: 97.50%**,
- test set size: 25,250 images.

These metrics confirm that ViT performs strongly on food imagery even under limited compute constraints.

B. Confusion Analysis

Misclassified samples were analyzed to identify common confusion patterns. Fig. 1 presents the most frequent confusion pairs.

Frequent errors arise from:

1) visually similar categories (e.g., steak vs. filet mignon),
2) bias toward the most salient ingredient rather than the full dish,
3) multi-item scenes with ambiguous dominant labels.

C. Nutrition Output Demonstration

For each prediction, the system retrieves a nutrient profile from the lookup table. Although serving size and recipe variation are not estimated, this step illustrates the feasibility of linking image recognition to nutritional interpretation.

IV. Conclusion and Future Work

A. Conclusion

This project presents an end-to-end pipeline for nutrition-aware food classification using the Food-101 dataset and a fine-tuned Vision Transformer. The model achieves strong quantitative performance, with 87% Top-1 and 97.5% Top-5 accuracy, and qualitative analyses highlight both strengths and common failure cases. Together, these results demonstrate the feasibility of using transformer-based architectures as a foundation for automated dietary assessment systems.

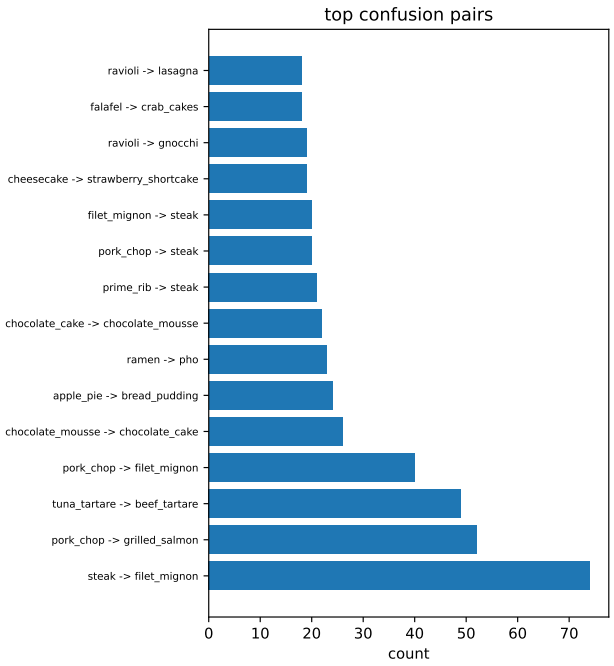

Fig. 1. Most frequent confusion pairs on the Food-101 test set.

B. Limitations

Despite promising performance, several limitations constrain real-world applicability:

- **Lack of portion estimation:** Images do not reliably convey portion size, ingredients, or cooking methods, all of which affect true nutritional content.
- **Coarse category definitions:** Food-101 labels group together visually similar but nutritionally distinct variants (e.g., types of pizza or salads).
- **Dataset bias:** Real-world photos differ from the curated images in Food-101, weakening generalization under domain shift.
- **Single-label design:** Many images contain multiple foods, yet only one label is provided.

C. Future Work

Future extensions could include:

1) fine-grained food classification with sub-category labels,
2) portion-size estimation via depth cues or reference objects,
3) domain adaptation using real-world mobile photos,
4) baseline and ablation studies for deeper model analysis,
5) multi-label classification for multi-food images,
6) user studies to evaluate practical utility and usability.

D. Final Remarks

Overall, this work demonstrates a practical and interpretable approach to bridging computer vision and nutrition estimation. By outlining limitations and proposing concrete future directions, the project contributes toward developing more robust, scalable, and user-centered diet-tracking tools.

REFERENCES

[1] L. Bossard, M. Guillaumin, and L. Van Gool, "Food-101: Mining discriminative components with random forests," *ECCV*, 2014.

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CVPR*, 2016.

[3] A. Dosovitskiy *et al.*, "An image is worth 16×16 words," *ICLR*, 2021.

[4] U.S. Department of Agriculture, "FoodData Central," https://fdc.nal.usda.gov.

[5] Hugging Face Documentation, https://huggingface.co/docs.